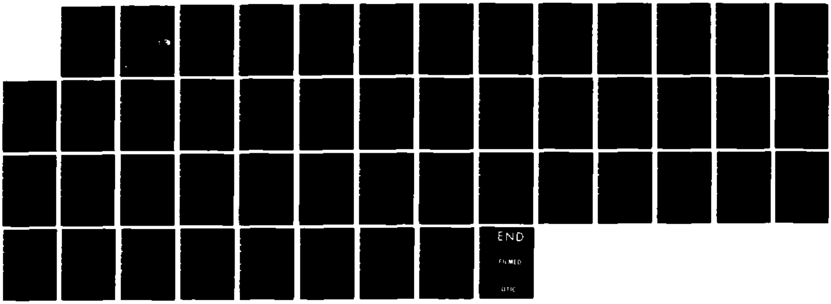
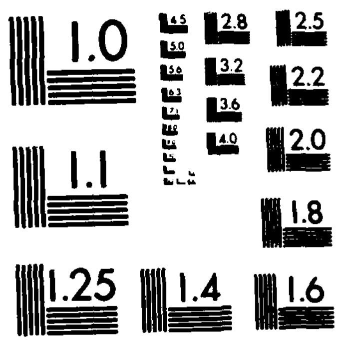
AD-A154 129 PERFORMANCE ENVELOPES AND OPTIMAL APPROPRIATENESS MEASUREMENT(U) ILLINOIS UNIV AT URBANA MODEL BASED MEASUREMENT LAB M V LEVINE ET AL. DEC 84 1/1

UNCLASSIFIED MEASUREMENT-SER-84-5 N00014-79-C-0752 F/G 5/10 NL





UNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE

AD-A154129

REPORT DOCUMENTATION PAGE

1a. REPORT SECURITY CLASSIFICATION
Unclassified

1b. RESTRICTIVE MARKINGS

2a. SECURITY CLASSIFICATION AUTHORITY

2b. DECLASSIFICATION / DOWNGRADING SCHEDULE

3. DISTRIBUTION / AVAILABILITY OF REPORT
Approved for public release:
distribution unlimited

4. PERFORMING ORGANIZATION REPORT NUMBER(S)
Measurement Series 84-5

5. MONITORING ORGANIZATION REPORT NUMBER(S)

6a. NAME OF PERFORMING ORGANIZATION
Michael V. Levine
Model-Based Measurement Lab.

6b. OFFICE SYMBOL *(If applicable)*

7a. NAME OF MONITORING ORGANIZATION
Personnel and Training Research Programs
Office of Naval Research

6c. ADDRESS *(City, State, and ZIP Code)*
University of Illinois
210 Education Bldg., 1310 S. Sixth St.
Champaign, IL 61820

7b. ADDRESS *(City, State, and ZIP Code)*
Code 442PT
800 North Quincy Street
Arlington, VA 22217

8a. NAME OF FUNDING / SPONSORING ORGANIZATION

8b. OFFICE SYMBOL *(If applicable)*

9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER

8c. ADDRESS *(City, State, and ZIP Code)*

10. SOURCE OF FUNDING NUMBERS

PROGRAM ELEMENT NO	PROJECT NO	TASK NO	WORK UNIT ACCESSION NO
61153N	RR042-04	RR042-04-01	NR 154-445 NR 150-518

11. TITLE *(Include Security Classification)*
Performance Envelopes and Optimal Appropriateness Measurement (unclassified)

12. PERSONAL AUTHOR(S)
Levine, Michael V. and Drasgow, Fritz

13a. TYPE OF REPORT
Technical Report

13b. TIME COVERED
FROM ________ TO ________

14. DATE OF REPORT *(Year, Month, Day)*
1984, December

15. PAGE COUNT
48

16. SUPPLEMENTARY NOTATION

17. COSATI CODES

FIELD	GROUP	SUB-GROUP

18. SUBJECT TERMS *(Continue on reverse if necessary and identify by block number)*
Latent trait theory, item response theory, multiple choice test, appropriateness measurement, person fit, appropriateness index, optimal test, symmetric functions, (cont.)

19. ABSTRACT *(Continue on reverse if necessary and identify by block number)*
The test-taking behavior of some examinees may be so idiosyncratic that their test scores are not comparable to the scores of more typical examinees. Appropriateness indices provide quantitative measures of response-pattern atypicality. An appropriateness index can be viewed as a test statistic for testing a null hypothesis of normal test-taking behavior against an alternative hypothesis of atypical test-taking behavior. In this paper performance curves and the performance envelope are introduced as devices for obtaining a least upper bound for the power of the most powerful statistical tests for aberrance. The performance envelope of a set of tests is the function on [0,1] whose value at t is the least upper bound of the hit rates of the tests when their false positive rate is t . The performance curve of an appropriateness index is the performance envelope of the tests for aberrance based on the index. For some types of testing anomalies it is possible to determine the performance envelope for the set of all statistical tests for aberrance and to identify a test whose performance curve is identical to this performance envelope. An algorithm for computing some of these optimal tests is described, and an example of its use is presented.

20. DISTRIBUTION / AVAILABILITY OF ABSTRACT
☐ UNCLASSIFIED/UNLIMITED ☒ SAME AS RPT ☐ DTIC USERS

21. ABSTRACT SECURITY CLASSIFICATION
unclassified

22a. NAME OF RESPONSIBLE INDIVIDUAL
Michael V. Levine

22b. TELEPHONE *(Include Area Code)*
217/333-0092

22c. OFFICE SYMBOL

DD FORM 1473, 84 MAR
83 APR edition may be used until exhausted
All other editions are obsolete

SECURITY CLASSIFICATION OF THIS PAGE
UNCLASSIFIED

18. Subject terms (continued)

Neyman-Pearson Lemma, hypothesis testing, cheating, copying, bias, deliberate failure.

Abstract

The test-taking behavior of some examinees may be so idiosyncratic that their test scores are not comparable to the scores of more typical examinees. Appropriateness indices provide quantitative measures of response-pattern atypicality. An appropriateness index can be viewed as a test statistic for testing a null hypothesis of normal test-taking behavior against an alternative hypothesis of atypical test-taking behavior. In this paper performance curves and the performance envelope are introduced as devices for obtaining a least upper bound for the power of the most powerful statistical tests for aberrance. The performance envelope of a set of tests is the function on [0,1] whose value at t is the least upper bound of the hit rates of the tests when their false positive rate is t . The performance curve of an appropriateness index is the performance envelope of the tests for aberrance based on the index. For some types of testing anomalies it is possible to determine the performance envelope for the set of all statistical tests for aberrance and to identify a test whose performance curve is identical to this performance envelope. An algorithm for computing some of these optimal tests is described, and an example of its use is presented.

Additional keywords: latent trait theory; item response theory; multiple choice tests; cheating; copying.

PERFORMANCE ENVELOPES

1. Introduction

An examinee's test-taking behavior may be so idiosyncratic that his/her test score is not comparable to the scores of more typical examinees. Copying and other forms of cheating could result in a spuriously high score. Language problems, atypical education and deliberate failure could result in a spuriously low score.

Some atypical examinees produce recognizably unusual answer patterns. For example, in a recent experimental study of deliberate failure, deliberately failing examinees often chose obviously incorrect options, whereas truly failing examinees rarely chose these options. Furthermore, deliberately failing examinees produced the option response sequence ADADAD relatively often; however, truly failing examinees very rarely produced this sequence.

Appropriateness measurement attempts to detect faulty test scores by recognizing unusual answer patterns. The standard procedure is to formulate a model for normal data and a model for aberrant data. With these models the identification of faulty test patterns is reduced to a hypothesis testing problem.

To date, appropriateness measurement studies have been highly empirical. For example, to determine if a form of aberrance can be detected, several plausible detection procedures are tried out on actual and simulation data containing normal and aberrant response patterns.

There are a number of questions that cannot be answered satisfactorily by these empirical methods. To return to the example, if none of the

evaluated detection procedures classifies well, then it cannot be concluded that the form of inappropriateness could not be detected because some other procedure may have worked well.

This paper introduces a general method for obtaining a less ambiguous answer to the question of whether a specific form of aberrance is detectable. Section Two presents some motivating examples of applications of our results. Section Three introduces terminology and some basic concepts. Section Four develops the basic theory and relates it to several important measurement questions. Section Five reviews two specific applications. Section Six contains some mathematical results that demonstrate that the theory can be implemented on currently available computers for these two applications. An algorithm for computing some performance envelopes is described in Section Seven. Section Eight provides an illustrative example.

2. Examples

In this section some examples are used to motivate and to describe our results.

Example One: Absolute detectibility of a specific form of aberrance in simulation data.

Simulation data sets are commonly used to compare tests for appropriateness and decide whether a specific form of appropriateness can be detected. These studies use a pair of computer programs, one to simulate normal response data and another to simulate patterns of right and wrong answers from aberrant examinees. Since an arbitrarily large number of examinees can be simulated, the performance of any statistical test for appropriateness can be accurately determined. This is done by computing the hit rate (proportion of aberrant examinees correctly classified) and false positive rate (proportion of normal examinees incorrectly classified) with large samples. If one finds a test with a high hit rate and low false positive rate, then one concludes the specified type of aberrance can be detected, at least in simulation data. (Of course, follow-up studies with actual data are needed to verify the simulation results. However, some of our results are more easily understood with simulation studies.)

Without loss of generality it can be assumed that the collection of statistical tests being considered contains at least one test with false alarm rate equal to α for every α between zero and one. It makes sense to determine the hit rate β of the most powerful test among those with a given false alarm rate. In fact it is possible to consider the set of all statistical tests and find a bound at each α . For some important special

cases we have developed a useable way to compute a least upper bound for β at each α. In fact is is possible to describe (and compute) a statistical test that actually achieves the maximum.

These results are important because, at least for simulation data, they yield an absolute measure of the detectibility of the specified form of appropriateness. Thus, after applying our methods to a particular appropriateness measurement problem one may be led to one of the following conclusions:

1. The specified form of aberrance cannot be detected very well by any appropriateness measurement technique, whatsoever; or
2. There is no point in attempting to improve upon a developed, convenient appropriateness measurement test because it is only slightly less powerful than all superior tests; or
3. There are tests that are substantially more powerful than the tests currently being used, and significant gains in power may be obtained by revising appropriateness measurement techniques.

Example Two: Choosing between dichotomous and polychotomous models.

Polychotomous analyses are considerably more difficult than analyses of multiple choice data scored right or wrong. For a specified form of aberrance, a specified population, and a specified multiple choice test, can one substantially improve appropriateness measurement procedures by attending to which wrong answer was chosen? The results in this paper are useful for answering at least some forms of this question.

Using the results in this paper, for any false alarm rate, the hit rate of the most powerful statistical test can be computed, at least for some polychotomous tests. The maximum is taken over all tests, including those

that are sensitive to which wrong answer was chosen. The maximum can also be computed for all tests that treat examinees with the same pattern of correct answers equally, i.e. for dichotomous statistical tests. By comparing maxima one can better decide if polychotomous analyses deliver enough additional power to be worth developing and implementing. If the maxima are close, then the increased sampling error in the polychotomous model's parameter estimates may off-set gains in statistical power.

Example Three: Descriptive models of actual data.

Since a multiple choice test has only finitely many items a Markovian model of high enough order will exactly describe the statistical structure of sampled examinees. Unless there are complex interdependencies between nonadjacent items, lower order Markovian models will adequately approximate the higher order, perfectly descriptive model. There are other families of models that also provide increasingly accurate and, finally, a perfectly descriptive model (e.g., Bahadur, 1968). The descriptive models generally require very large samples for parameter estimation; however, in some appropriateness measurement tasks, very large samples of normal and aberrant examinee data are available for parameter estimation.

In a recent study of deliberate test failure, Markovian models of order n_T and n_E were fitted to large samples of truly failing examinees and experimental examinees deliberately failing an exam. For each pair of models, the reasoning used in this paper was applied to compute an optimal test for inappropriateness in Markovian data. For $n_T=1$ and $n_E=2$ a test was obtained which, for actual data, was clearly more powerful than all available alternative appropriateness tests.

To summarize, the results in this paper can be applied to a sequence of models of increasing generality and used to approximate a bound on the performance of an optimal test for aberrance. In the process of approximating a bound, a powerful test for aberrance will be constructed.

3. Ability Distributions, Sampling and Optimal Tests

The typical problem for appropriateness measurement is to find a statistical test δ such that $\delta(u)$ is much more likely to indicate aberrance when the response vector **u** has been generated by a sampled aberrant examinee than when u has been generated by a sampled normal examinee. The key word in this description of appropriateness measurement is "sampled." From a practical point of view, it makes sense to consider examinees as sampled since they report for testing in a haphazard order and since, except when they are cheating, they work independently of one another. From a theoretical point of view it is desirable to regard examinees as sampled because doing so leads to multinomial item response models, simple (as opposed to composite) statistical hypotheses, and optimal appropriateness measurement tests. A brief discussion of item response theory will clarify these points.

The equations of item response theory are consistent with many conflicting psychological interpretations. The most useful one for appropriateness measurement, in our opinion, is to regard each examinee as having an ability θ and item scores $u_1, u_2, \ldots, u_n$. The item scores, according to this view, are random variables because the set of all examinees is a probability space and not because any examinee's behavior is uncertain. Similarly, θ is a random variable only in the sense that probabilities are assigned to sets of examinees with specified θ values. "Measurement error" is irrelevant to θ's status as a random variable. Thus for any examinee, say examinee ω, $u_1(\omega)$ and $\theta(\omega)$ are numbers indicating ω's response and ability, respectively. The probability that u_1 is zero or that θ is negative, on the other hand, are the

probabilities assigned to the set of examinees answering the first item incorrectly or having an ability less than zero. Thus if examinees are regarded as sampled, $\text{Prob}\{u_1=0\}$ and $\text{Prob}\{\theta<0\}$ are the probabilities of sampling an examinee with an incorrect first answer and negative ability.

For reference, the defining equations of item response theory are reproduced below. Our results follow from these equations and do not depend upon viewing subjects as deterministic and sampled.

The basic assumption of item response theory, the local independence assumption, is generally formulated with reference to the item response functions, $P_1(\cdot), P_2(\cdot), \ldots, P_n(\cdot)$, which give the conditional probabilities of correct $(u_i=1)$ responses at each ability level. Local independence asserts that

$$(3.1) \qquad \text{Prob}\{u_1=u_1^* \;\&\; u_2=u_2^* \ldots \&\; u_n=u_n^* \mid \theta=t\} = \prod_{j=1}^{n} P_i(t)^{u_i^*}[1-P_i(t)]^{1-u_i^*} .$$

where u_i^*, the observed item score, is either zero or one.

When the ability density is known or accurately estimated, equation (3.1) can be used to compute unconditional probabilities. If the ability random variable has density f, then the probability that the response vector $\mathbf{u}$ equals some vector of zeros and ones $\mathbf{u^*}$ is obtained by integrating the likelihood function

$$(3.2) \qquad \text{Prob}\{\mathbf{u}=\mathbf{u}^*\} = \int \text{Prob}\{\mathbf{u}=\mathbf{u}^* \mid \theta=t\} f(t)dt .$$

In many item response theory applications the ability density is ignored. When an ability density is not specified, then the likelihood function (3.1) specifies a continuum of models for normal item response

data, one for each ability. The hypothesis, " $\mathbf{u^*}$ has been generated by a normal examinee," is composite in the sense that there is a different probability that $\mathbf{u=u^*}$ for each ability level t . Such a formulation leads to maximum likelihood ratio tests such as Levine and Rubin's (1979) LR test.

When the ability density can be specified or accurately estimated, then the hypothesis that $\mathbf{u^*}$ has been generated by a normal examinee is simple in the sense that formula (3.2) gives a unique model consistent with the hypothesis. When the alternative hypothesis is also simple, then the likelihood ratio can be used to obtain an optimal test for appropriateness. According to the Neyman-Pearson Lemma (Lehmann, 1959) a statistical test of the form

$$\delta(\mathbf{u^*}) = \begin{cases} \text{"aberrant," if } P_{\text{Aberrant}}(\mathbf{u^*}) \geq \text{constant} \cdot P_{\text{Normal}}(\mathbf{u^*}) \\ \text{"normal," otherwise} \end{cases}$$

has as much or more power for detecting aberrance than all tests with the same false positive rate.

Note that when the ability density is specified, item response data are multinomial with 2^n categories. Multinomial conceptualizations of the usual models for aberrant data will be formulated as they are needed. The key point of this section is that classical statistical results for testing simple hypotheses can be used without making implausible psychometric assumptions.

4. Performance Envelopes

Each of the examples in Section Two was concerned with a set of statistical tests. For example, the second example compared a set of tests using polychotomous data with a set of tests that can be applied to dichotomous data. In this section a device for studying properties of sets of tests, the performance envelope, is introduced. But first some notation and terminology are needed.

The basic data for appropriateness measurement are the vectors of item responses, here denoted by $\mathbf{u}$. A deterministic or nonrandomized statistical test for aberrance is a binary function of item responses taking on the values 1 (to indicate aberrance) and 0 (to indicate the absence of aberrance). Following Lehmann (1959, p. 60), a pair of tests can be combined to form a randomized test. If $\delta_1(u)$ and $\delta_2(u)$ are tests and $0 \leq p \leq 1$ then $d(u;\delta_1,\delta_2,p)$ is used to denote the randomized test which is $\delta_1(u)$ with probability p and $\delta_2(u)$ with probability $1-p$.

This paper is exclusively concerned with properties of sets of statistical test of aberrance, such as the set of all tests that can be obtained from a given goodness-of-fit statistic or the set of all statistics that can be obtained using a given model for test data. The mathematics of comparing sets of tests is simplest when these sets are closed with respect to routine operations and methods for combining tests.

If $\mathbf{D}$ is a set of statistical tests, then a set $\bar{\mathbf{D}}$, possibly equal to $\mathbf{D}$, is defined as the set of tests obtainable from tests in $\mathbf{D}$ by "probability mixtures" (i.e., forming randomized tests from pairs, triples or larger finite sets of tests), complementation (i.e., forming the test $1-\delta$ from δ), and considering the trivial test (i.e. the test $\delta_0(u)=1$,

which labels all patterns as aberrant). If no new tests can be constructed by these routine operations on the tests of **D** , i.e., if $\mathbf{D}=\overline{\mathbf{D}}$, then **D** will be called closed. In most cases of interest (see below), explicitly expressing all the tests of $\overline{\mathbf{D}}$ with formulas containing only the tests of **D** is straightforward.

To evaluate the performance of a (randomized or nonrandomized) test for aberrance, two conditional probabilities are needed. Using the suggestive terminology of signal detection theory, these are the false positive rate $\alpha(\delta)$ or probability of misclassifying a randomly sampled normal examinee and the hit rate $\beta(\delta)$ or probability of correctly classifying a sampled aberrant examinee. In hypothesis testing terminology, these are the probability of a type I error and the power of δ respectively.

Of course a pair of distributions $P_{\text{Aberrant}}(u)$ and $P_{\text{Normal}}(u)$ over response vectors must be specified to make the phrases "randomly sampled normal examinee" and "sampled aberrant examinee" unambiguous. For each individual application this will be done.

To evaluate the performance of the most powerful tests that can be obtained from a set of tests **D** , a monotonic real function is introduced, the performance envelope. If **D** is a set of statistical tests, then the performance envelope of **D** is the function $R=R_{\mathbf{D}}$ defined for $0 \le t \le 1$ by

$$R(t) = \text{least upper bound } \{\beta(\delta):\ \delta \in \overline{\mathbf{D}} \text{ and } \alpha(\delta)=t\} .$$

It is easy to prove R is a non-decreasing function with values between zero and one.

Two special cases, the performance curve of a statistic and the performance envelope for the set of all statistical tests, will now be used to illustrate the definition.

4.1 The Performance Curve for a Statistic

Let X be a test statistic, i.e., a number-valued function of item responses such as any of the many goodness-of-fit indicators proposed as an index of appropriateness. For each "critical score" c, two statistical tests for aberrance can be formulated. One of them

$$\delta_c(u) = \begin{cases} 1, \text{ if } & X(\mathbf{u}) \leq c \\ 0, \text{ if } & X(\mathbf{u}) > c \end{cases}$$

treats low values of X as indicative of aberrance, and the other, $1-\delta_c$, treats high values as indicative of aberrance. The <u>performance curve for the statistic X</u> is the performance envelope of the set of statistics of form δ_c or $1-\delta_c$.

The performance curve of X is important because it shows how well X performs in classifying examinees at each false alarm rate, in the following sense. Let D_X denote the set of all tests of form δ_c. For each t, there will be a statistical test δ obtainable from D_X with false alarm rate equal to t and hit rate equal to $R_{D_X}(t)$. This test can be regarded as most powerful or optimal among the tests obtainable from D_X with $\alpha=t$ because every other test (with false alarm rate equal to t) will have lower or equal hit rate. The word "obtainable" seems especially apt here because it is easy to show[1] that the optimal test can always be chosen to be one of the nonrandomized tests or a randomized test obtained from just two nonrandomized tests.

The performance curve for X differs from the ROC curve for X usually used in appropriateness measurement in that it is continuous and concave. (Recall that the ROC curve for X is the set of points $\langle x,y \rangle$

with $x=\alpha(\delta_c)$ and $y=\beta(\delta_c)$ for some nonrandomized test obtainable from X. Some authors use "ROC curve" to denote a curve obtained by fitting a linear or other smooth curve between points and thus obtaining a continuous, but not necessarily concave function.)

Since there are only finitely many response patterns, there are only finitely many points $\langle\alpha(\delta_c),\beta(\delta_c)\rangle$. If the piecewise linear curve obtained by connecting points corresponding to consecutive values of c is the graph of a concave function, then this function is the performance curve for X.

Computation of the performance curve for X becomes slightly more complicated if the ROC has points below the diagonal or if the curve obtained by connecting consecutive points is not concave. One considers the finite set of points $\langle\alpha,\beta\rangle$ obtained from all the non-randomized tests. One obtains a curve as a piecewise linear function beginning with the origin (or the point with highest β from among all those with $\alpha=0$ in case there are nontrivial tests with $\alpha=0$) as the first node. If $\langle\alpha,\beta\rangle$ is the n^{th} node of the piecewise linear function, then the next node is $\langle\alpha',\beta'\rangle$ where α',β' maximize $(\beta'-\beta)/(\alpha'-\alpha)$ over the subset of the finite set with $\alpha'>\alpha$.

The performance curve is preferable to the ROC for comparing a pair of statistics X and Y for two reasons. First, for each t either $R_X(t)\geq R_Y(t)$ or $R_X(t)<R_Y(t)$ so the choice between X and Y is clear when a false alarm rate t is desired, even when there is no nonrandomized test with false alarm rate t. Second, the performance curve for a statistic X is concave, but the ROC curve need not be. Thus for some range of possible false alarm rates α, say between $t-\varepsilon$ and $t+\varepsilon$ for $\varepsilon>0$, a randomized test can have higher hit rate than all the nonrandomized

tests δ with $t-\varepsilon \leq \alpha(\delta) \leq t+\varepsilon$. Consequently comparing ROC curves can lead to the wrong choice between X and Y to use for constructing a statistical test with false alarm rate near t .

In concluding this subsection we wish to point out that sets of tests like D_X are much more general than seems to be realized. A set of nonrandomized statistical tests for aberrance has nested critical regions if for any δ_1 and δ_2 in D either $\delta_1 \leq \delta_2$ or $\delta_2 \leq \delta_1$. In other words, $\delta_1(u^*) \leq \delta_2(u^*)$ for every response pattern u^* or $\delta_2(u^*) \leq \delta_1(u^*)$ for every response pattern u^* . Using the fact that there are only finitely many possible response patterns it can be shown that if D has nested critical regions there is a statistic X such that $\bar{D}=D_X$ and the performance envelope for D is the performance curve for X .

This fact is important because it shows the generality of the approach to appropriateness measurement we use: classifying examinees by using an "appropriateness index" or real valued function of item scores and a range of cutting scores. Any set of tests with nested critical regions can be obtained with this approach.

4.2 The Performance Envelope for the Set of All Statistical Tests

At least in some situations it is practical to consider the performance envelope for the set of all statistical tests for aberrance, and this leads to a second illustration of performance envelopes.

As noted in Section Three when the ability distribution is specified formula (3.2) defines a simple multinomial model for item response patterns. Plausible, simple multinomial models (e.g. the spurious high and spurious low models of Sections Five and Six) are appropriate for some important

forms of aberrant data. Thus in principal the likelihood ratio statistic

$$\lambda(u) = P_{Aberrant}(u)/P_{Normal}(u)$$

can be defined. In Section Seven an algorithm for calculating λ is described.

A basic result for this research is that the performance envelope for the set of all statistical tests for aberrance is the performance curve for the likelihood ratio statistic. In other words, for any statistical test δ , there is a test obtainable from the likelihood ratio statistic with false alarm rate equal to $\alpha(\delta)$ and hit rate at least as large as $\beta(\delta)$. This fundamental result is an immediate consequence of the Neyman-Pearson Lemma (Lehmann, 1959, p. 63).

5. Spurious Scores and the Computation of Envelopes

In the remainder of this paper an algorithm for computing performance envelopes for some important models is described and illustrated. The spurious score models and tampering manipulations have provided reference experiments for comparing appropriateness measurement results in several laboratories (Drasgow, Levine, & Williams, 1985; Levine & Rubin, 1979; Parsons, 1983; Rudner, 1983). Spurious score model and tampering experiments are also important because they can be used to predict the performance of appropriateness measurement procedures in various actual situations without collecting additional data.

The 10% spurious high tampering manipulation is an operation on an actual or simulation examinee's answer sheet that changes up to 10% of the examinee's item scores. In this manipulation 10% of the items are sampled without replacement. Incorrect answers are changed to correct answers, and correct answers are left unchanged.

Data conform to a 10% spurious high model if the likelihood function for each item response pattern is the likelihood function for a response pattern generated by a normal examinee and then modified by 10% spurious high tampering. An explicit formula is given later in this section.

The spurious high model and tampering procedures were formulated after considering a low ability examinee copying from a much brighter neighbor when the proctor happened to be distracted. Of course, some copiers will risk copying on 10% of the items and others on 20% or 5% of the items. However, after a distribution on the percentages is specified, results from studies in which the percent tampering has been constant can be combined to approximate performance in the more realistic situation. The studies in

which percent tampering is constant are basic because they permit the psychometrician to predict for an arbitrary percent copying distribution.

According to the 10% spurious high model, the likelihood of an item response pattern $\mathbf{u}^* = (u_1^*, u_2^*, \ldots, u_n^*)$ is a sum over $\binom{n}{n/10}$ terms

$$\binom{n}{n/10}^{-1} \sum_S \prod_{i \in S} u_i^* \prod_{i \notin S} P_i(t)^{u_i^*} Q_i(t)^{1-u_i^*}$$

where S ranges over subsets of the first n positive integers having exactly $n/10$ elements. Direct computation of likelihoods is impractical because for $n=90$, $n/10=9$ there are more than 10^{11} terms.

The 10% <u>spurious low tampering manipulation</u> is a procedure that also revises normal item response patterns. Exactly 10% of the examinee's item responses are sampled. For each sampled item response a random response is generated. If the generated response agrees with the examinee's response, no change is made. Otherwise, the examinee's item response is changed to the generated response. <u>Spurious low score models</u> are defined analogously to spurious high score models by referring to a two stage procedure; the first stage conforms to a model for normal responding, and the second stage modifies the patterns generated in the first stage by a spurious low tampering manipulation. The likelihood function for this model is

$$\binom{n}{n/10}^{-1} \sum_S \prod_{i \in S} A_i^{u_i^*}(1-A_i)^{1-u_i^*} \prod_{i \notin S} P_i(t)^{u_i^*} Q(t)^{1-u_i^*}$$

where the summation is over subsets of the first n positive integers having exactly $n/10$ elements and where the A_i are taken to be one over the number of options for item i .

The spuriously high model models copiers and examinees with knowledge of a test's answer key for some proportion of the test items. The

spuriously low model models random responding to some proportion of test items.

Spurious low aberrance can also be interpreted in meaningful ways. Consider, for example, the assessment of children for possible assignment to special education programs. There are serious concerns about the meanings of test scores when tests standardized on mainstream samples are administered to cultural minorities. This is particularly important when a child is tested in a second language in which he or she may not be fluent. His or her responses to some linguistically demanding items may be nearly random. The seriousness of this problem is underscored by the fact that "intelligence" tests cannot be used in California when assessing minority children for special education (see Hulin, et al., 1983, Chapter 9). As before, results with fixed percentages of tampering can be combined to predict for situations in which the number of spurious items has an arbitrary specified distribution.

6. An Algorithm for Calculating Likelihoods

The major obstacle to computing performance envelopes for the spurious models is the calculation of likelihoods. An algorithm for computing these likelihoods can be obtained from classical results on symmetric functions. In this section a highly intuitive derivation not requiring symmetric functions is given. The intuitive derivation has the advantage of showing that the algorithm can be used to study a large variety of appropriateness problems. It appears useful for modeling tests in which items differ in the degree to which they elicit an aberrant response and in which there are complex interactions between ability and tendency to cheat or otherwise perform aberrantly.

Consider an experiment in which on each trial an examinee is presented an item. Suppose on trial i the examinee performs normally with probability $1-\rho$ but responds aberrantly with probability ρ so that the probability of a correct response can be written

$$[1-\rho_i(t,s)]P_i(t) + \rho_i(t,s)A_i(t) .$$

For example an examinee with an imperfect "crib sheet," ability t and inclination to cheat s risks using the crib sheet to answer item i with probability $\rho_i(t,s)$ and then answers correctly with probability $A_i(t)$ or chooses to ignore the crib sheet witn probability $1-\rho_i(t,s)$ and then answers correctly with probability $P_i(t)$. In this interpretation of the equations, $A_i(t) = 1$ if the crib sheet has the correct answer, zero if the crib sheet has the wrong answer and $P_i(t)$ if the crib sheet has no information on the item. In our analyses of spurious high and low models,

$A_i(t)$ will be 1 or the reciprocal of the number of options, and ρ will also be independent of i, t and s.

If the appropriate independence assumptions are made, the likelihood function for a response pattern $\mathbf{u}^*$ will be

$$\ell(\mathbf{u}^*;t,s) = \prod_{i=1}^{n} \{[1-\rho_i(t,s)]P_i(t) + \rho_i(t,s)A_i(t)\}^{u_i^*} \times \{[1-\rho_i(t,s)]Q_i(t) + \rho_i(t,s)[1-A_i(t)]\}^{1-u_i^*}$$

If $\rho_i(t,0)=0$, then $\ell(\mathbf{u}^*;t,0)$ is the likelihood function for normal examinees.

In many analyses it is necessary to keep track of the number of items on which cheating or aberrant responding took place. To this end an indeterminate r is introduced and a "probability generating function" G is defined by

$$G(\mathbf{u}^*;r,t,s) = \prod_{i=1}^{n} \{[1-\rho_i(t,s)]P_i(t) + r\rho_i(t,s)A_i(t)\}^{u_i^*} \times \{[1-\rho_i(t,s)]Q_i(t) + r\rho_i(t,s)[1-A_i(t)]\}^{1-u_i^*}$$

If G is written as a polynomial in r, then the constant term, $G(\mathbf{u}^*;0,t,s)$, is the probability of observing $\mathbf{u}^*$ from an examinee making no aberrant responses. The linear term, $\frac{\partial}{\partial r} G\big|_{r=0}$, is the probability of observing $\mathbf{u}^*$ from examinees making exactly one aberrant response. More generally, the coefficient of r^k (i.e. $(1/k!)\frac{\partial^k}{\partial r^k} G$ evaluated $r=0$) will be the probability of pattern $\mathbf{u}^*$ with exactly k aberrant responses. If $\rho_i(t,s) = .5$ for all i then the coefficient of r^k is $.5^n$ times the sum of the products having exactly k factors selected from the set

$\{A_i(t)^{u_i^*}[1-A_i(t)]^{1-u_i^*}: i=1, \ldots, n\}$ and $n-k$ factors from $\{P_i(t)^{u_i^*} Q_i^{1-u_i^*}: i=1, \ldots, n\}$. In other words, the coefficient of r^k is $\binom{n}{k}$ (i.e., the number of ways to select k items from n) times $.5^n$ times the probability of observing **u*** when exactly k responses are aberrant and all the subsets of k responses are equally likely.

To simplify the evaluation of these coefficients G is divided by the constant term to obtain

$$\frac{G(\mathbf{u^*},r,s,t)}{G(\mathbf{u^*},0,s,t)} = \prod_{i=1}^{n} [1+rB_i] ,$$

$$\text{where} \quad B_i = \begin{cases} \dfrac{\rho_i(t,s)}{[1-\rho_i(t,s)]} \times \dfrac{A_i(t)}{P_i(t)}, & \text{if } u_i^* = 1 , \\[2ex] \dfrac{\rho_i(t,s)}{[1-\rho_i(t,s)]} \times \dfrac{1-A_i(t)}{Q_i(t)}, & \text{if } u_i^* = 0 . \end{cases}$$

Note that if $\rho_i(t,s)$ equals $.5$ for each item i, the terms in ρ drop out, and the coefficient of r^k in $\Pi[1+rB_i]$ is $\binom{n}{k}\ell(\mathbf{u^*};t,0)^{-1}$ times the probability of a $k/n \times 100\%$ percent spurious (high/low) examinee producing pattern **u***, provided the $A_i(t)$ terms are appropriately chosen.

This formula permits enormous computational savings because the coefficients of the powers of r can be computed recursively with relatively few operations. Since

$$\begin{aligned} \prod_{i=1}^{m} (1+rB_i) &= [1+rB_m] \prod_{i=1}^{m-1} (1+rB_i) \\ &= \prod_{i=1}^{m-1} (1+rB_i) + rB_m \prod_{i=1}^{m-1} (1+rB_i) \end{aligned}$$

it is clear that the coefficients in the partial products

$$\prod_{i=1}^{m} (1+rB_i) = C_{0,m} + rC_{1,m} + r^2C_{2,m} + \ldots$$

satisfy the recursion

$$C_{r,m+1} = C_{r,m} + B_{m+1}C_{r-1,m}$$

where $C_{0,m} = 1$ and $C_{r,m} = 0$ for $r>m$.

To illustrate the use of this formula consider 10.6% spurious low tampering on an 85 item test. The P_i are specified as three parameter logistic functions and $A_i(t)=.2$ was used to model a random choice from the five multiple choice options. The aberrant items were obtained by sampling 9 items from all 85 without replacement. The likelihood of a particular pattern **u*** being sampled from among all examinees having parameters t,s and producing exactly 9 aberrant responses is the sum of $\binom{85}{9} = 7.1 \times 10^{11}$ products, each of which has many factors. There is one product for each way to select 9 responses from 85. Thus a direct computation requires 85×10^{11} multiplications at each ability level.

By using the recursion the number of multiplications can be greatly reduced. The desired probability is equal to

$$\binom{85}{9}^{-1} \ell(\mathbf{u^*},t,0) \times C_{9,85}$$

where $C_{9,85}$ is the coefficient r^9 in the polynomial

$$\prod_{i=1}^{85} [1+rB_i] .$$

and where the B_i's are computed by setting $\rho_i(t,s) = 1/2$ and $A_i(t) = .2$.

To calculate $C_{9,85}$ a 10 entry array is revised 85 times. Initially C_0 is set equal to 1 , and the remaining C's , C_1, C_2, . . . , C_9 , are

set equal to zero. The m^{th} revision replaces C_r by the current value of C_r plus B_m times the current value of C_{r-1} for $r=1, 2, \ldots, 9$. C_0 is left equal to 1. Thus the eighty five revisions require less than 850 multiplications.

7. An Algorithm for Computing Some Performance Envelopes

In this section an algorithm is described for computing performance envelopes for the set of all statistical tests in the important situations in which

1. item response functions are specified
2. ability distributions are specified for both the normal and aberrant populations, and
3. data from aberrant examinees conform to a spurious high model or a spurious low model.

Each of these conditions is commented upon separately below.

1. Specified item response functions certainly pose no problem for the reference simulation studies that are commonly performed. A variety of item response function estimation procedures are available for actual data (Bock & Aitkin, 1981; Lord, 1968; Samejima, 1981). Levine and Drasgow (1982) reported experiments for measuring the effects of using estimated item response functions in appropriateness measurement studies with actual and simulated data and in which the item parameter estimation sample contains a specified proportion of unidentified aberrant examinees. They found that with sufficiently large item parameter estimation samples and parameters estimated with LOGIST (Wood, Wingersky, & Lord, 1976) from samples with and without aberrants the index L_0 performed about as well with estimated item parameters as with correct item parameters. Portions of their studies are currently being repeated to gauge the effects of using estimated parameters on performance envelopes.

2. Ability distribution estimation programs are available (e.g. Levine, 1984, 1985; Mislevy, 1984) for dealing with normal populations.

Levine has shown that his method is strongly consistent and asymptotically efficient (1985). Much less is known about estimating ability distribution for aberrant examinees. Furthermore, the aberrant sample will generally be quite small. However, sometimes it is acceptable to assume that ability has the same distribution in both populations; other times the ability distribution can be specified by apriori considerations. For example, one of the hardest and most important tasks for appropriateness measurement is to identify spuriously high cheaters with ability slightly below the minimum required to qualify for military technical training. To measure performance in this worst case, the aberrant distribution is assumed to uniform over a short interval below the critical ability.

3. In the example presented in the next section 10% spurious low aberrance is studied. Essentially the same algorithm is used for spurious high aberrance. We feel that the constant percentage spuriousness condition is especially important because, as noted in Section Five, these studies are used as reference experiments and because the constant percentage studies can be easily combined to predict performance without collecting new data after virtually any distribution over percent spuriousness has been chosen or estimated. However, by appropriately specifying the $\rho_i(t,s)$ and $A_i(t,s)$ in Section Six, item effects and complex interactions between ability and "inclination towards aberrance" can be modelled. For example two values of ρ_i could be used to model the fact that only some of the items were available to a coaching school or a dishonest military recruiter. The s variable could be used as an index when modelling second language problems in a population consisting of several distinct linguistic groups, say hispanics, Mandarin speaking Chinese Americans and examinees speaking

English only. In any event the basic algorithm suffices for a variety of optimal appropriateness measurement problems.

To obtain the performance envelope for the set of all statistics, the performance curve for the likelihood ratio statistic λ

$$\lambda(u^*) = P_{\text{Aberrant}}(u=u^*)/P_{\text{Normal}}(u=u^*)$$

is computed. To approximate the λ performance curve the sample λ ROC curve is calculated for a large sample normal and aberrant examinees. By using the fact that $\lambda(\mathbf{u})$ assumes only finitely many values it is easy to show that with probability one the piecewise linear function connecting consecutive points on the sample ROC converges to the performance curve for λ .

To calculate $\lambda(\mathbf{u}^*)$ the numerator and denominator are calculated separately. For normal examinees, the likelihood function is calculated by substituting the specified item parameters in

$$P_i(t) = c_i + \frac{1-c_i}{1+\exp[-a_i(t-b_i)]} ,$$

and numerically integrating as in equation (3.2) to obtain

$$P_{\text{Normal}}(u^*) = \int \prod_i \{[P_i(t)]^{u_i^*}[1-P_i(t)]^{1-u_i^*}\} f(t)dt .$$

P_{Aberrant} is also an integrated likelihood function. The computation of the integrand is discussed later in this section after f and the integration are described.

In our research to date, we have taken the density f to be normal (0,1) or normal (0,1) truncated to the interval [-2.05,2.05] when generating simulation data and evaluating the integrals to compute P_{Normal}

and $P_{Aberrant}$. Although normality is not required, our current algorithm does take advantage of some of its properties. In particular, it uses the facts that the normal density is continuous and "flat" relative to the likelihood functions for abilities less than 2.05 in absolute value. The normal density varies from .054 to .399 on the interval [-2.05,2.05] , but the likelihood function's maximum is usually 10^{10} to 10^{20} times as large as its minimum on the interval for the 85 to 95 item tests we have studied. Consequently, portions of the interval [-2.05,2.05] can often be ignored with little loss of accuracy when computing probabilities.

The integrals in P_{Normal} and $P_{Aberrant}$ are being evaluated by Simpson's rule. For both probabilities we obtained four to five digit accuracy when the distance Δ between quadrature points was .20 and five to six digit accuracy for $\Delta = .10$. We have generally used $\Delta = .10$ in our calculations because it seemed to provide the best trade-off between numerical accuracy and computing expense.

The number of function evaluations can be reduced by first computing the maximum likelihood estimate $\hat{\theta}$ of ability given u^* . Let g denote the function to be integrated. Then g can be evaluated at points $\hat{\theta}-i\Delta$, $i=1, 2, \ldots, m_1$, until $g(\hat{\theta}-i\Delta)$ becomes very small. The algorithm requires $g(\hat{\theta}-i\Delta)$ to be less than 10^{-4} times as large as $g(\hat{\theta})$. Similarly, g can be evaluated at points $\hat{\theta}+i\Delta$, $i=1, 2, \ldots, m_2$. If the total number of function evaluations is odd, then Simpson's rule can be applied immediately. When the total is even, one more function evaluation should be obtained before application of Simpson's rule. We have found that the number of function evaluation can often be reduced by 50% for $\Delta=.10$ by this rule.

The recursive algorithm described in Section Six is used to calculate the likelihood function for aberrant examinees. The algorithm is first summarized with no more generality than is needed for the spurious high and low studies. The remainder of this section discusses refinements of the basic algorithm for spurious high and low studies.

Recall that the likelihood function for spurious high aberrance is

$$P_A(u=u^*|\theta=t) = \sum_S \text{Probability \{set S is sampled for tampering\}} \times$$

$$\prod_{i=1}^{n} \text{Prob}\{u_i=u_i^*|\theta=t \text{ and S is sampled}\}$$

$$= \binom{n}{k}^{-1} \sum_S \prod_{i\in S} u_i^* \prod_{i\notin S} P_i(t)^{u_i^*} Q_i(t)^{1-u_i^*} .$$

Now if S contains one or more of the incorrectly answered items $\prod_{i\in S} u_i^* = 0$. Consequently the summation can be taken over all k element subsets of the n_c correctly answered items rather than of the n items, and the second product will be divisible by $W(t) = \prod_{i:u_i^*=0} Q_i(t)$. Thus

$$P_A(u=u^*|\theta=t) = \binom{n}{k}^{-1} W(t) \sum_S \prod_{\substack{i:i\notin S' \,\& \\ u_i^*=1}} P_i(t)$$

where the summation is over the $\binom{n_c}{k}$ k-element subsets S' of the set of correct items in pattern $\mathbf{u^*}$. In other words, the summation is the (n_c-k)th symmetric function on the vector of n_c not necessarily distinct variables $\langle P_{i_1}(t), P_{i_2}(t), \ldots, P_{i_{n_c}}(t)\rangle$ where $i_j < i_{j+1}$ and $u_{i_j}^* = 1$. To evaluate the summation we use the well-known recursion given a probabilistic interpretation in Section Six

$$T(r+1,j) = T(r,j) + x_{r+1}T(r,j-1)$$

discussed in Section Six. Here $T(i,j)$ is the i^{th} elementary symmetric function on the first j variables in a vector or sequence of numbers $\langle x_1, x_2, \ldots \rangle$, i.e. the sum of the $\binom{j}{i}$ products having i factors selected from the first j numbers.

For spurious low aberrance the likelihood function is

$$P_{Aberrant}(u=u^*|\theta=t) = \sum_S \text{Probability \{set S is sampled for tampering\}} \times$$
$$\prod_{i=1}^{n} \text{Prob}\{u_i=u_i^*|\theta=t \text{ and S is sampled}\}$$
$$= \binom{n}{k}^{-1} \sum_S \prod_{i\epsilon S} p^{u_i^*}(1-p)^{1-u_i^*} \prod_{i \notin S} P_i(t)^{u_i^*} Q_i(t)^{1-u_i^*} .$$

where the summation is over k element subsets S of the n items and where $p=.2$ is the probability of being correct when responding randomly on a 5 option multiple choice test. To expeditiously calculate the likelihood for a pattern $\mathbf{u^*}$ with n_c correct and $n_w=n-n_c$ wrong we rewrite this as

$$\binom{n}{k}^{-1} p^{n_c}(1-p)^{n_w} \sum_S \prod_{i \notin S} [P_i(t)/p]^{u_i^*}[Q_i(t)/(1-p)]^{1-u_i^*}$$

and evaluate $\sum_S \prod_{i \notin S} [P_i(t)/p]^{u_i^*}[Q_i(t)/(1-p)]^{1-u_i^*}$ as the $(n-k)^{th}$ symmetric function on the vector $\langle [P_1(t)/p]^{u_1^*}[Q_1(t)/(1-p)]^{1-u_1^*}, \ldots, [P_n(t)/p]^{u_n^*}[Q_n(t)/(1-p)]^{1-u_n^*} \rangle$

A considerable further reduction in computation can be obtained by using the fact that the $(m-k)^{th}$ symmetric function in $\langle x_1, \ldots, x_m \rangle$ equals $\prod_{i=1}^{m} x_i$ times the k^{th} symmetric function in $\langle x_1^{-1}, \ldots, x_m^{-1} \rangle$.

Thus

$$\sum_S \prod_{i \notin S} [P_i(t)/p]^{u_i^*}[Q_i(t)/(1-P)]^{1-u_i^*}$$

$$= \prod_{i=1}^{n} [P_i(t)/p]^{u_i^*}[Q_i(t)/(1-p)]^{1-u_i^*} \times$$

$$\sum_S \prod_{i \in S} [p/P_i(t)]^{u_i^*}[q/Q_i(t)]^{1-u_i^*}$$

$$= p^{-n_c} q^{-n_w} \ell(\mathbf{u^*},t) \times \text{ the } k^{th} \text{ symmetric function in}$$

$$\langle [p/P_1(t)]^{u_1^*}[q/Q_1(t)]^{1-u_1^*}, \ldots, [p/P_n(t)]^{u_n^*}[q/Q_n(t)]^{1-u_n^*} \rangle$$

The same identity gives a reduction in the amount of calculation for spurious high analyses for patterns $\mathbf{u^*}$ with $k < n_c - k$.

8. An Illustrative Example

To illustrate the algorithm described in Section Seven, item parameter estimates obtained from Levine and Drasgow's (1983) fitting of the three parameter logistic model to the 85 item April, 1975 Scholastic Aptitude Test-Verbal section (SAT-V) were taken as simulation parameters. One thousand normal response vectors were created by sampling abilities from a normal (0,1) distribution truncated to the [-2.05,2.05] interval, computing the logistic probabilities of correct responses, and then scoring each simulated item response as correct or incorrect depending upon whether a number sampled from a uniform [0,1] distribution was less than or greater than the logistic probability. A sample of 500 spuriously low response patterns was created by first generating 500 normal response patterns. Then nine simulated items were randomly selected without replacement from each response pattern and each item was rescored to be correct with probability .2 and rescored to be incorrect with probability .8 . The likelihood ratio statistic was computed for all 1500 patterns, as described in Section Seven.

Table One presents the proportions of spuriously low response patterns correctly classified as aberrant when various proportions of normal response patterns are misclassified as aberrant. The table also presents the results for the standardized ℓ_o index studied by Drasgow, Levine, and Williams (1985). It is evident that the envelope curve statistic provides a substantial improvement over the standardized ℓ_o index. This finding is important because in previous research (Drasgow, 1982; Levine & Drasgow, 1982; Levine & Rubin, 1979) we have been unable to find an index that provides detection rates that are clearly superior to ℓ_o .

Table 1

Proportions of Spuriously Low Responses Patterns

Classified as aberant at Various False Alarm Rates

Proportion of Normal Response Patterns Classified As Aberrant	Proportion Detected by Envelope Curve Statistic	Proportion Detected by Standardized ℓ_o	Hit Rate Ratio
.005	.114	.060	.526
.010	.132	.070	.530
.015	.144	.096	.667
.020	.170	.112	.659
.030	.198	.142	.717
.040	.228	.182	.798
.050	.276	.210	.761
.060	.294	.250	.850
.080	.328	.286	.872
.100	.368	.322	.875
.150	.452	.390	.863
.200	.532	.452	.850
.250	.590	.486	.824
.300	.634	.554	.874
.400	.734	.666	.907
.500	.804	.742	.923

Footnotes

This work was supported by United States Office of Naval Research contracts N00014-79C-0752, NR 154-445 and N00014-83K-0397, NR 150-518, Michael V. Levine, Principal Investigator.

[1] $\langle t, R_{D_X}(t)\rangle$ is on the boundary of a convex polygon because the range of X is finite. Therefore $\langle t, R_{D_X}(t)\rangle$ is a vertex (and a nonrandomized test is optimal) or $\langle t, R_{D_X}(t)\rangle$ is on a line segment connecting two vertices (and a randomized test obtained from the two tests associated with the segment is optimal).

REFERENCES

Bahadur, R.R. (1968). A representation of the joint distribution of responses to _n_ dichotomous items. In H. Solomon, _Studies in item analysis and prediction_. Stanford, California: Stanford University Press.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F.M. Lord & M.R. Novick, _Statistical theories of mental test scores_. Reading, Mass.: Addison-Wesley.

Bock, R.D. & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. _Psychometrika_, _46_, 443-459.

Drasgow, F. (1982). Choice of test model for appropriateness measurement. _Applied Psychological Measurement_, _6_, 297-308.

Drasgow, F., Levine, M.V., & Williams, E. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. _British Journal of Mathematical and Statistical Psychology_, in press.

Hulin, C.L., Drasgow, F., & Parson, C.K. (1983). _Item response theory: Application to psychological measurement_. Homewood, Ill.: Dow Jones-Irwin.

Lehmann, E.L. (1959). _Testing statistical hypotheses_. New York: Wiley.

Levine, M.V. (1985). Representing ability distributions. Report 85-1. Champaign, IL: Model-Based Measurement Laboratory, Department of Educational Psychology, University of Illinois.

Levine, M.V. (1984). _An introduction to multilinear formula score theory_. Champaign, IL: Model-Based Measurement Laboratory, Department of Educational Psychology, University of Illinois.

Levine, M.V. & Drasgow, F. (1982). Appropriateness measurement: Review, critique and validating studies. _British Journal of Mathematical and Statistical Psychology_, _35_, 42-56.

Levine, M.V. & Drasgow, F. (1983). The relation between incorrect option choice and estimated ability. _Educational and Psychological Measurement_, _43_, 675-685.

Levine, M.V. & Rubin, D.F. (1979). Measuring the appropriateness of multiple choice test scores. _Journal of Educational Statistics_, _4_, 269-290.

Lord, F.M. (1968). An analysis of the Verbal Scholastic Aptitude Test using Birnbaum's three-parameter logistic model. _Educational and Psychological Measurement_, _28_, 989-1020.

Mislevy, R.J. (1984). Estimating latent distributions. Psychometrika, 49, 359-382.

Parsons, C.K. (1983). The identification of people for whom JDI scores are inappropriate. Organizational Behavior and Human Performance, 31, 365-393.

Rudner, L.M. (1983). Individual assessment accuracy. Journal of Educational Measurement, 20, 207-219.

Samejima, F. (1981). Final report: Efficient methods of estimating the operating characteristics of item response categories and challenge to a new model for the multiple-choice item. Technical Report. Knoxville, Tennessee: Department of Psychology, University of Tennessee.

Wood, R.L., Wingersky, M.S., & Lord, F.M. (1976). LOGIST - A computer program for estimating examinee ability and item characteristic curve parameters. Research Memorandum 76-6. Princeton, N.J.: Educational Testing Service.

Distribution List

Personnel Analysis Division
AF/MPXA
5C360, The Pentagon
Washington, DC 20330

Air Force Human Resources Lab
AFHRL/MPD
Brooks AFB, TX 78235

Air Force Office
of Scientific Research
Life Sciences Directorate
Bolling Air Force Base
Washington, DC 20332

Dr. Robert Ahlers
Code N711
Human Factors Laboratory
NAVTRAEQUIPCEN
Orlando, FL 32813

Dr. Erling B. Andersen
Department of Statistics
Studiestraede 6
1455 Copenhagen
DENMARK

Technical Director
Army Research Institute for the
Behavioral and Social Sciences
5001 Eisenhower Avenue
Alexandria, VA 22333

Special Assistant for Projects
OASN(M&RA)
5D800, The Pentagon
Washington, DC 20350

Dr. Alan Baddeley
Medical Research Council
Applied Psychology Unit
15 Chaucer Road
Cambridge CB2 2EF
ENGLAND

Dr. Patricia Baggett
University of Colorado
Department of Psychology
Boulder, CO 80309

Dr. Isaac Bejar
Educational Testing Service
Princeton, NJ 08450

CDR Robert J. Biersner, USN
Naval Biodynamics Laboratory
P. O. Box 29407
New Orleans, LA 70189

Dr. Menucha Birenbaum
School of Education
Tel Aviv University
Tel Aviv, Ramat Aviv 69978
Israel

Dr. Werner Birke
Personalstammamt
der Bundeswehr
D-5000 Koeln 90
WEST GERMANY

Code N711
Attn: Arthur S. Blaiwes
Naval Training Equipment Center
Orlando, FL 32813

Dr. R. Darrell Bock
University of Chicago
Department of Education
Chicago, IL 60637

Dr. Nick Bond
Office of Naval Research
Liaison Office, Far East
APO San Francisco, CA 96503

Dr. Robert Breaux
Code N 095R
NAVTRAEQUIPCEN
Orlando, FL 32813

Dr. Robert Brennan
American College Testing
Programs
P. O. Box 168
Iowa City, IA 52243

Dr. Patricia A. Butler
NIE Mail Stop 1806
1200 19th St., NW
Washington, DC 20208

Dr. James Carlson
American College Testing
Program
P.O. Box 168
Iowa City, IA 52243

Dr. John B. Carroll
409 Elliott Rd.
Chapel Hill, NC 27514

Dr. Robert Carroll
NAVOP 01B7
Washington, DC 20370

Mr. Raymond E. Christal
AFHRL/MOE
Brooks AFB, TX 78235

Dr. Norman Cliff
Department of Psychology
Univ. of So. California
University Park
Los Angeles, CA 90007

Director
Manpower Support and
Readiness Program
Center for Naval Analysis
2000 North Beauregard Street
Alexandria, VA 22311

Scientific Advisor
to the DCNO (MPT)
Center for Naval Analysis
2000 North Beauregard Street
Alexandria, VA 22311

Chief of Naval Education
and Training
Liason Office
AFHRL
Operations Training Division
Williams AFB, AZ 85224

Assistant Chief of Staff
Research, Development,
Test, and Evaluation
Naval Education and
Training Command (N-5)
NAS Pensacola, FL 32508

Office of the Chief
of Naval Operations
Research Development
& Studies Branch
NAVOP 01B7
Washington, DC 20350

Dr. Stanley Collyer
Office of Naval Technology
800 N. Quincy Street
Arlington, VA 22217

Dr. Hans Crombag
University of Leyden
Education Research Center
Boerhaavelaan 2
2334 EN Leyden
The NETHERLANDS

CTB/McGraw-Hill Library
2500 Garden Road
Monterey, CA 93940

CDR Mike Curran
Office of Naval Research
800 N. Quincy St.
Code 270
Arlington, VA 22217-5000

Mr. Timothy Davey
University of Illinois
Educational Psychology
Urbana, IL 61801

Dr. Dattprasad Divgi
Syracuse University
Department of Psychology
Syracuse, NY 13210

Dr. Hei-Ki Dong
Ball Foundation
800 Roosevelt Road
Building C, Suite 206
Glen Ellyn, IL 60137

Dr. Fritz Drasgow
University of Illinois
Department of Psychology
603 E. Daniel St.
Champaign, IL 61820

Defense Technical
Information Center
Cameron Station, Bldg 5
Alexandria, VA 22314
Attn: TC
(12 Copies)

Dr. Stephen Dunbar
Lindquist Center
for Measurement
University of Iowa
Iowa City, IA 52242

Dr. Kent Eaton
Army Research Institute
5001 Eisenhower Blvd.
Alexandria, VA 22333

Dr. John M. Eddins
University of Illinois
252 Engineering Research
Laboratory
103 South Mathews Street
Urbana, IL 61801

Dr. Susan Embertson
University of Kansas
Psychology Department
Lawrence, KS 66045

ERIC Facility-Acquisitions
4833 Rugby Avenue
Bethesda, MD 20014

Dr. Benjamin A. Fairbank
Performance Metrics, Inc.
5825 Callaghan
Suite 225
San Antonio, TX 78228

Dr. Pat Federico
Code P13
NPRDC
San Diego, CA 92152

Dr. Leonard Feldt
Lindquist Center
for Measurment
University of Iowa
Iowa City, IA 52242

Dr. Richard L. Ferguson
American College Testing
Program
P.O. Box 168
Iowa City, IA 52240

Dr. Gerhard Fischer
Liebiggasse 5/3
A 1010 Vienna
AUSTRIA

Dr. Myron Fischl
Army Research Institute
5001 Eisenhower Avenue
Alexandria, VA 22333

Prof. Donald Fitzgerald
University of New England
Department of Psychology
Armidale, New South Wales 2351
AUSTRALIA

Mr. Paul Foley
Navy Personnel R&D Center
San Diego, CA 92152

Dr. Bob Frey
Commandant (G-P-1/2)
USCG HQ
Washington, DC 20593

Dr. Janice Gifford
University of Massachusetts
School of Education
Amherst, MA 01002

Dr. Robert Glaser
Learning Research
& Development Center
University of Pittsburgh
3939 O'Hara Street
Pittsburgh, PA 15260

Dr. Bert Green
Johns Hopkins University
Department of Psychology
Charles & 34th Street
Baltimore, MD 21218

H. William Greenup
Education Advisor (E031)
Education Center, MCDEC
Quantico, VA 22134

Dipl. Pad. Michael W. Habon
Universitat Dusseldorf
Erziehungswissenshaftliches
Universitatsstr. 1
D-4000 Dusseldorf 1
WEST GERMANY

Dr. Ron Hambleton
School of Education
University of Massachusetts
Amherst, MA 01002

Prof. Lutz F. Hornke
Universitat Dusseldorf
Erziehungswissenschaftliches
Universitatsstr. 1
Dusseldorf 1
WEST GERMANY

Dr. Paul Horst
677 G Street, #184
Chula Vista, CA 90010

Mr. Dick Hoshaw
NAVOP-135
Arlington Annex
Room 2834
Washington, DC 20350

Dr. Lloyd Humphreys
University of Illinois
Department of Psychology
603 East Daniel Street
Champaign, IL 61820

Dr. Steven Hunka
Department of Education
University of Alberta
Edmonton, Alberta
CANADA

Dr. Earl Hunt
Department of Psychology
University of Washington
Seattle, WA 98105

Dr. Huynh Huynh
College of Education
Univ. of South Carolina
Columbia, SC 29208

Dr. Douglas H. Jones
Advanced Statistical
Technologies Corporation
10 Trafalgar Court
Lawrenceville, NJ 08148

Prof. John A. Keats
Department of Psychology
University of Newcastle
N.S.W. 2308
AUSTRALIA

Dr. Norman J. Kerr
Chief of Naval Education
and Training
Code 00A2
Naval Air Station
Pensacola, FL 32508

Dr. William Koch
University of Texas-Austin
Measurement and Evaluation
Center
Austin, TX 78703

Dr. Leonard Kroeker
Navy Personnel R&D Center
San Diego, CA 92152

Dr. Patrick Kyllonen
AFHRL/MOE
Brooks AFB, TX 78235

Dr. Anita Lancaster
Accession Policy
OASD/MI&L/MP&FM/AP
Pentagon
Washington, DC 20301

Dr. Daryll Lang
Navy Personnel R&D Center
San Diego, CA 92152

Dr. Jerry Lehnus
OASD (M&RA)
Washington, DC 20301

Dr. Thomas Leonard
University of Wisconsin
Department of Statistics
1210 West Dayton Street
Madison, WI 53705

Dr. Alan M. Lesgold
Learning R&D Center
University of Pittsburgh
Pittsburgh, PA 15260

Dr. Michael Levine
Educational Psychology
210 Education Bldg.
University of Illinois
Champaign, IL 61801

Dr. Charles Lewis
Faculteit Sociale Wetenschappen
Rijksuniversiteit Groningen
Oude Boteringestraat 23
9712GC Groningen
The NETHERLANDS

Dr. Robert Linn
College of Education
University of Illinois
Urbana, IL 61801

Dr. Robert Lockman
Center for Naval Analysis
200 North Beauregard St.
Alexandria, VA 22311

Dr. Frederic M. Lord
Educational Testing Service
Princeton, NJ 08541

Dr. James Lumsden
Department of Psychology
University of Western Australia
Nedlands W.A. 6009
AUSTRALIA

Dr. William L. Maloy (02)
Chief of Naval Education
and Training
Naval Air Station
Pensacola, FL 32508

Dr. Gary Marco
Stop 31-E
Educational Testing Service
Princeton, NJ 08451

Dr. Clessen Martin
Army Research Institute
5001 Eisenhower Blvd.
Alexandria, VA 22333

Dr. Scott Maxwell
Department of Psychology
University of Notre Dame
Notre Dame, IN 46556

Dr. Samuel T. Mayo
Loyola University of Chicago
820 North Michigan Avenue
Chicago, IL 60611

Dr. James McBride
Psychological Corporation
c/o Harcourt, Brace,
Javanovich Inc.
1250 West 6th Street
San Diego, CA 92101

Dr. Clarence McCormick
HQ, MEPCOM
MEPCT-P
2500 Green Bay Road
North Chicago, IL 60064

Dr. Barbara Means
Human Resources
Research Organization
1100 South Washington
Alexandria, VA 22314

Dr. Robert Mislevy
Educational Testing Service
Princeton, NJ 08541

Dr William Montague
NPRDC Code 13
San Diego, CA 92152

Ms. Kathleen Moreno
Navy Personnel R&D Center
Code 62
San Diego, CA 92152

Headquarters, Marine Corps
Code MPI-20
Washington, DC 20380

Director
Research & Analysis Division
Navy Recruiting Command (Code 22)
4015 Wilson Blvd.
Arlington, VA 22203

Program Manager for Manpower,
Personnel, and Training
NAVMAT 0722
Arlington, VA 22217-5000

Dr. W. Alan Nicewander
University of Oklahoma
Department of Psychology
Oklahoma City, OK 73069

Dr. William E. Nordbrock
FMC-ADCO Box 25
APO, NY 09710

Dr. Melvin R. Novick
356 Lindquist Center
for Measurement
University of Iowa
Iowa City, IA 52242

Director, Manpower and Personnel
Laboratory
NPRDC (Code 06)
San Diego, CA 92152

Library
Code P201L
Navy Personnel R&D Center
San Diego, CA 92152

Technical Director
Navy Personnel R&D Center
San Diego, CA 92152

Commanding Officer
Naval Research Laboratory
Code 2627
Washington, DC 20390

Dr. Harry F. O'Neil, Jr.
Training Research Lab
Army Research Institute
5001 Eisenhower Avenue
Alexandria, VA 22333

Dr. James Olson
WICAT, Inc.
1875 South State Street
Orem, UT 84057

Mathematics Group
Office of Naval Research
Code 744MA
800 North Quincy Street
Arlington, VA 22217-5000

Office of Naval Research
Code 442PT
800 N. Quincy Street
Arlington, VA 22217-5000
(5 Copies)

Special Assistant for Marine
Corps Matters
Code 100M
Office of Naval Research
800 N. Quincy St.
Arlington, VA 22217-5000

Commanding Officer
Army Research Institute
ATTN: PERI-BR (Dr. J. Orasanu)
5001 Eisenhower Avenue
Alexandria, VA 22333

Dr. Jesse Orlansky
Institute for Defense Analyses
1801 N. Beauregard St.
Alexandria, VA 22311

Dr. Randolph Park
AFHRL/MOAN
Brooks AFB, TX 78235

Wayne M. Patience
American Council on Education
GED Testing Service, Suite 20
One Dupont Cirle, NW
Washington, DC 20036

Dr. James Paulson
Department of Psychology
Portland State University
P.O. Box 751
Portland, OR 97207

Dr. Roger Pennell
Air Force Human Resources
Laboratory
Lowry AFB, CO 80230

Administrative Sciences Department
Naval Postgraduate School
Monterey, CA 93940

Department of Operations Research
Naval Postgraduate School
Monterey, CA 93940

Dr. Mark D. Reckase
ACT
P. O. Box 168
Iowa City, IA 52243

Dr. Malcolm Ree
AFHRL/MP
Brooks AFB, TX 78235

Dr. Carl Ross
CNET-PDCD
Building 90
Great Lakes NTC, IL 60088

Mr. Robert Ross
Army Research Institute
5001 Eisenhower Avenue
Alexandria, VA 22333

Dr. Lawrence Rudner
403 Elm Avenue
Takoma Park, MD 20012

Dr. J. Ryan
Department of Education
University of South Carolina
Columbia, SC 29208

Dr. Fumiko Samejima
Department of Psychology
University of Tennessee
Knoxville, TN 37916

Mr. Drew Sands
NPRDC Code 62
San Diego, CA 92152

Dr. Robert Sasmor
Army Research Institute
5001 Eisenhower Avenue
Alexandria, VA 22333

Lowell Schoer
Psychological & Quantitative Foundations
College of Education
University of Iowa
Iowa City, IA 52242

Dr. Mary Schratz
Navy Personnel R&D Center
San Diego, CA 92152

Dr. W. Steve Sellman
OASD(MRA&L)
2B269 The Pentagon
Washington, DC 20301

Dr. Sylvia A. S. Shafto
National Institute of Education
1200 19th Street
Mail Stop 1806
Washington, DC 20208

Dr. Joyce Shields
Army Research Institute
5001 Eisenhower Avenue
Alexandria, VA 22333

Dr. Kazuo Shigemasu
7-9-24 Kugenuma-Kaigan
Fujusawa 251
JAPAN

Dr. William Sims
Center for Naval Analysis
200 North Beauregard Street
Alexandria, VA 22311

Dr. H. Wallace Sinaiko
Manpower Research and Advisory Services
Smithsonian Institution
801 North Pitt Street
Alexandria, VA 22314

Dr. Richard Snow
Liaison Scientist
Office of Naval Research Branch Office, London
Box 39
FPO New York, NY 09510

Dr. Richard Sorensen
Navy Personnel R&D Center
San Diego, CA 92152

Dr. Paul Speckman
University of Missouri
Department of Statistics
Columbia, MO 65201

Martha Stocking
Educational Testing Service
Princeton, NJ 08541

Dr. Peter Stoloff
Center for Naval Analysis
200 North Beauregard Street
Alexandria, VA 22311

Dr. William Stout
University of Illinois
Department of Mathematics
Urbana, IL 61801

Maj. Bill Strickland
AF/MPXOA
4E168 Pentagon
Washington, DC 20330

Dr. Hariharan Swaminathan
Laboratory of Psychometric and
Evaluation Research
School of Education
University of Massachusetts
Amherst, MA 01003

Mr. Brad Sympson
Navy Personnel R&D Center
San Diego, CA 92152

Dr. John Tangney
AFOSR/NL
Bolling AFB, DC 20332

Dr. Kikumi Tatsuoka
CERL
252 Engineering Research
Laboratory
Urbana, IL 61801

Dr. Maurice Tatsuoka
220 Education Bldg
1310 S. Sixth St.
Champaign, IL 61820

Dr. David Thissen
Department of Psychology
University of Kansas
Lawrence, KS 66044

Mr. Gary Thomasson
University of Illinois
Educational Psychology
Champaign, IL 61820

Dr. Robert Tsutakawa
Department of Statistics
University of Missouri
Columbia, MO 65201

Dr. Ledyard Tucker
University of Illinois
Department of Psychology
603 E. Daniel Street
Champaign, IL 61820

Dr. Vern W. Urry
Personnel R&D Center
Office of Personnel Management
1900 E. Street, NW
Washington, DC 20415

Dr. David Vale
Assessment Systems Corp.
2233 University Avenue
Suite 310
St. Paul, MN 55114

Dr. Frank Vicino
Navy Personnel R&D Center
San Diego, CA 92152

Dr. Howard Wainer
Division of Psychological Studies
Educational Testing Service
Princeton, NJ 08540

Dr. Ming-Mei Wang
Lindquist Center
for Measurement
University of Iowa
Iowa City, IA 52242

Mr. Thomas A. Warm
Coast Guard Institute
P. O. Substation 18
Oklahoma City, OK 73169

Dr. Brian Waters
HumRRO
300 North Washington
Alexandria, VA 22314

Dr. Edward Wegman
Office of Naval Research
Code 411
800 North Quincy Street
Arlington, VA 22217-5000

Dr. David J. Weiss
N660 Elliott Hall
University of Minnesota
75 E. River Road
Minneapolis, MN 55455

Dr. Donald Weitzman
MITRE
1820 Dolley Madison Blvd.
MacLean, VA 22102

Major John Welsh
AFHRL/MOAN
Brooks AFB, TX 78223

Dr. Douglas Wetzel
Code 12
Navy Personnel R&D Center
San Diego, CA 92152

Dr. Rand R. Wilcox
University of Southern California
Department of Psychology
Los Angeles, CA 90007

German Military Representative
ATTN: Wolfgang Wildegrube
Streitkraefteamt
D-5300 Bonn 2
4000 Brandywine Street, NW
Washington, DC 20016

Dr. Bruce Williams
Department of Educational Psychology
University of Illinois
Urbana, IL 61801

Dr. Hilda Wing
Army Research Institute
5001 Eisenhower Ave.
Alexandria, VA 22333

Ms. Marilyn Wingersky
Educational Testing Service
Princeton, NJ 08541

Dr. Martin F. Wiskoff
Navy Personnel R & D Center
San Diego, CA 92152

Mr. John H. Wolfe
Navy Personnel R&D Center
San Diego, CA 92152

Dr. George Wong
Biostatistics Laboratory
Memorial Sloan-Kettering Cancer Center
1275 York Avenue
New York, NY 10021

Dr. Wallace Wulfeck, III
Navy Personnel R&D Center
San Diego, CA 92152

Dr. Wendy Yen
CTB/McGraw Hill
Del Monte Research Park
Monterey, CA 93940

Major Frank Yohannan, USMC
Headquarters, Marine Corps
(Code MPI-20)
Washington, DC 20380

Dr. Joseph L. Young
Memory & Cognitive Processes
National Science Foundation
Washington, DC 20550

END

FILMED

7-85

DTIC